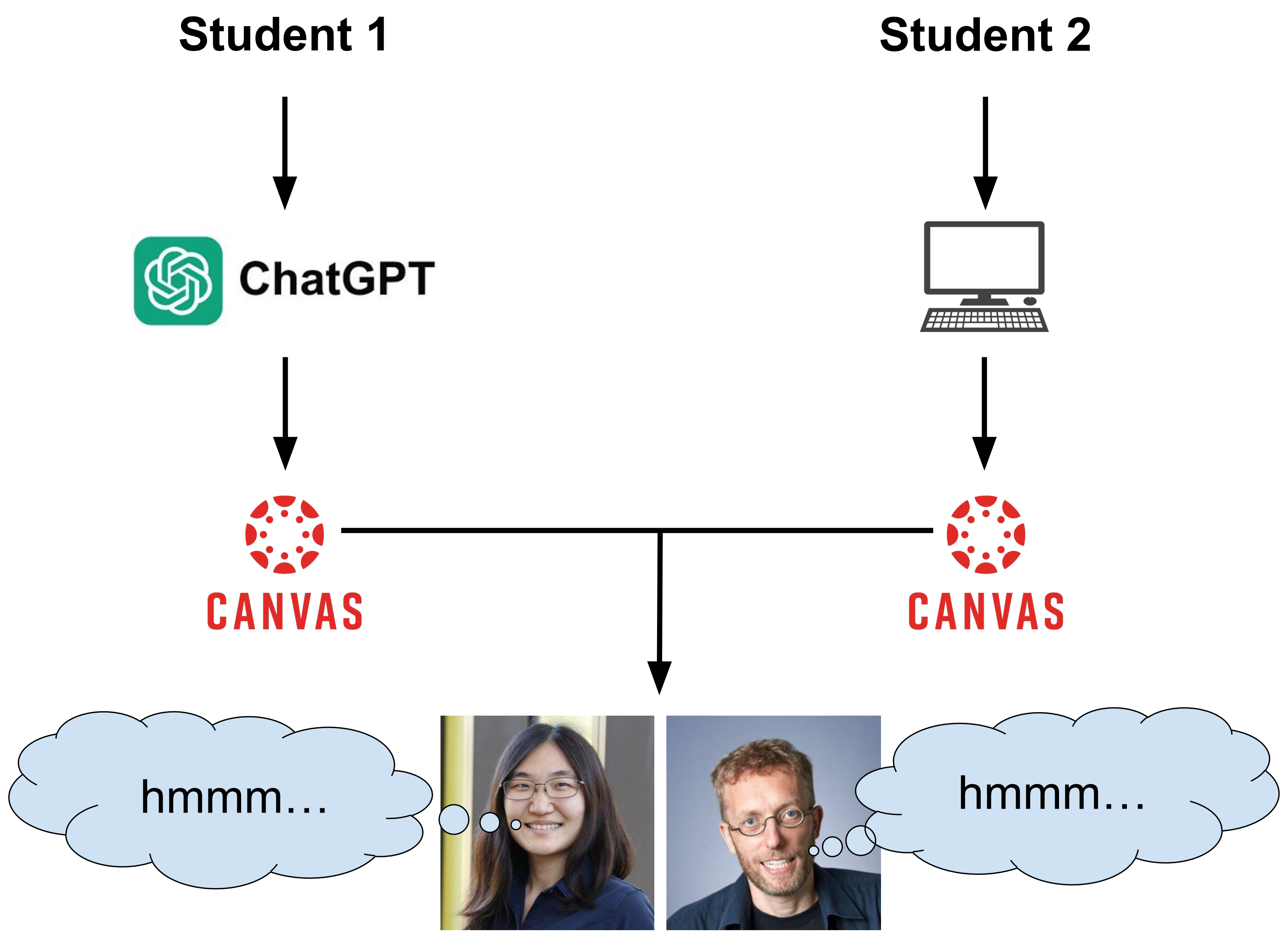
A Watermark for Large Language Models

CS 4782

Austin Houck, Tejasvi Yadav, Berk Gokmen, Arya Bhandari, Gemmechu Hassena

Introduction



Problem: How can we reliably distinguish between human-written and LLM-generated text?

Watermarking!

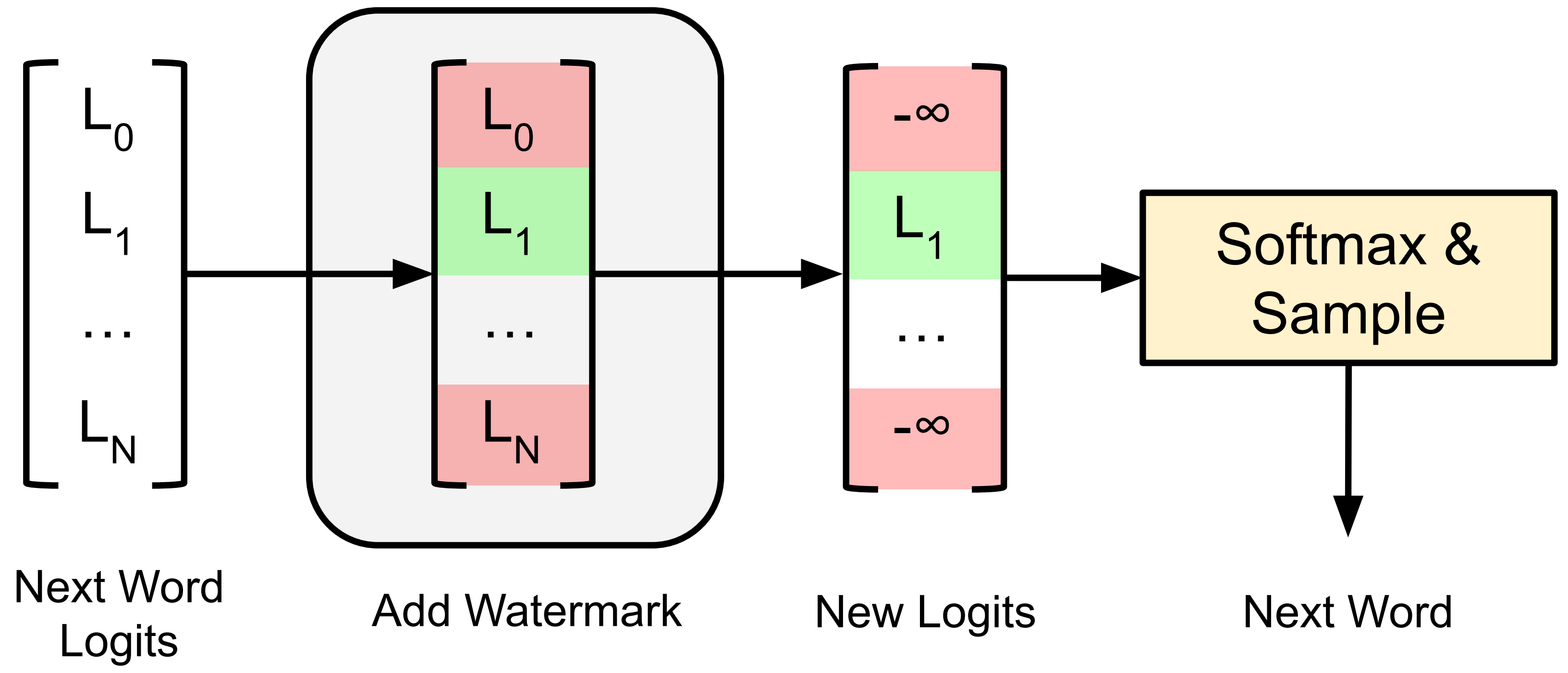
Goal: Add watermarking to any LLM output

- ➔ Without having to retrain.
- ➔ Without having to understand the architecture.
- ➔ That is compute efficient.
- ➔ That is easy to detect.

Methodology

- ➔ Used Meta's **OPT-1.3B** model for text generation.
- ➔ Employed prompts from the **RealNewsLike** subset of the **C4** dataset.
- ➔ Previous token's logits are used as the seed for randomness for reproducibility.
- ➔ Quantified watermark detectability through **z-test scoring**.

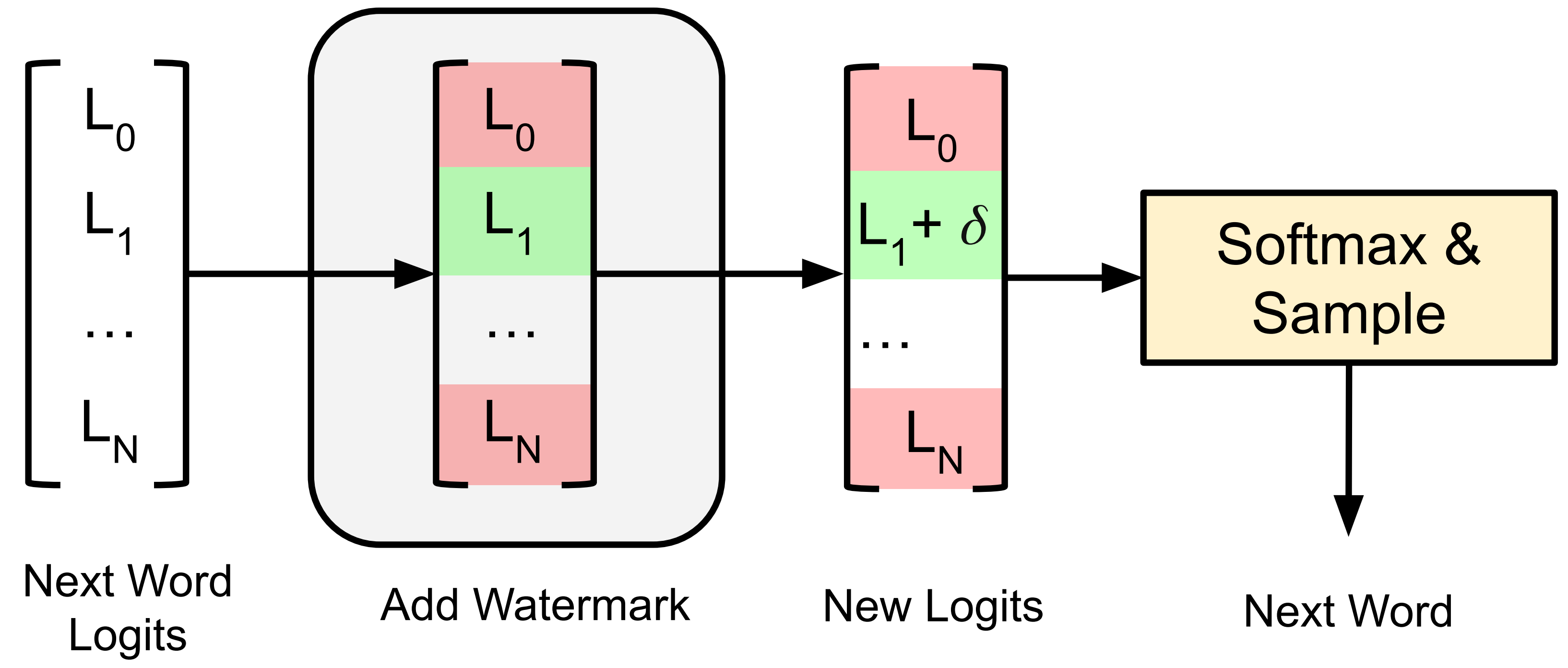
Hard Red List Watermarking



✅ Simple & extremely easy to detect.

❌ Text quality suffers in **low entropy** settings.

Soft Red List Watermarking

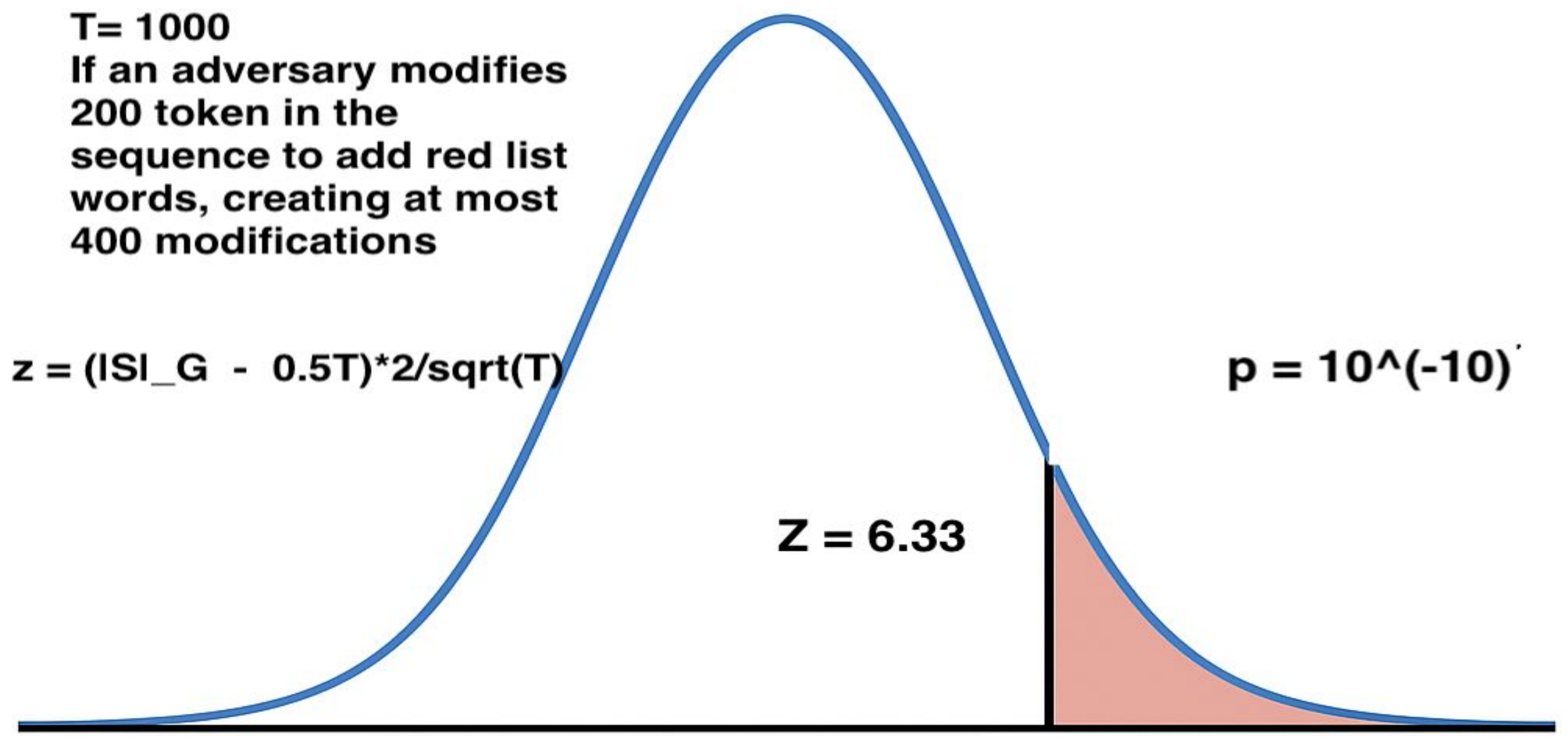


✅ Works well with **low entropy** settings

❌ May be harder to detect in certain settings

Attacks

H0: The text was generated with no knowledge of the red list rule



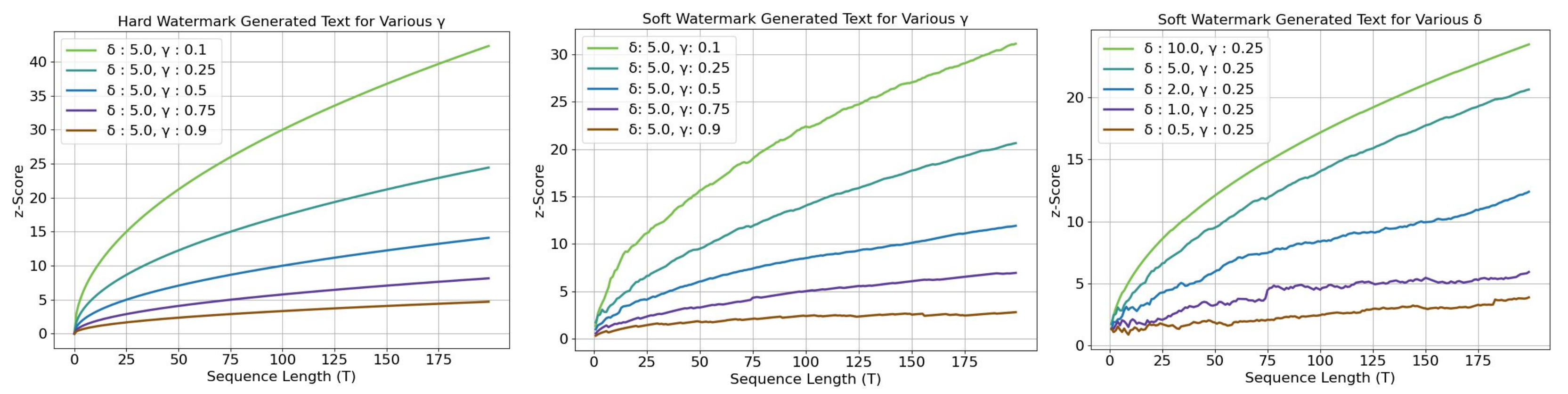
Probability distribution under the null hypothesis

Results

Prompt	No Watermark (NW)	Watermark (W)	↓ Entropy (S)	↑Z-Score (W)	↓PPL (NW)	↓PPL (W)
… While Arlington pored over the details, the Virginia General Assembly passed, and Democratic Gov. Ralph Northam signed, an incentives package worth up to $750 million for Amazon.	The package includes a $15 million cash grant, a 20 percent tax credit for Amazon's first five years of operations and $300 million in local property tax relief over the next decade…	Local officials said they were confident the deal would be approved.\nBut opponents of the deal have been waiting to hear the terms of the agreement…	0.84	8.68	8.47	8.48
… Services will include full security programs, compliance, third party vendor risk assessment, threat management, and managed security services.	\nKispert's appointment comes shortly after Grassi announced the retirement of Bruce Beyer, a 32-year veteran of the firm.	\n"I am excited to join the team at DMC and look forward to providing clients with the security and privacy services they need," Kispert said.	0.81	3.21	32.4	12.3
The vanishing gradient problem in Deep Learning is	a problem of finding a curve on a set of data points that represents the intensity of the point. The graph that the point should be in is the so called vanishing gradient.\n\nThe vanishing gradient problem is a very important one and it is one that has been studied extensively in the field of Deep Learning for decades…	… we are going to explore the application area of the vanishing gradient problem in deep learning by considering two different deep learning architectures, namely, the Linear Deep Learning Architecture and the Recurrent Deep Learning Architecture. We are also going to explore the application area of the vanishing gradient problem in deep learning by considering two different deep learning architectures, namely, the Linear Deep Learning Architecture and the Recurrent Deep Learning Architecture…	0.76	8.39	6.60	4.28

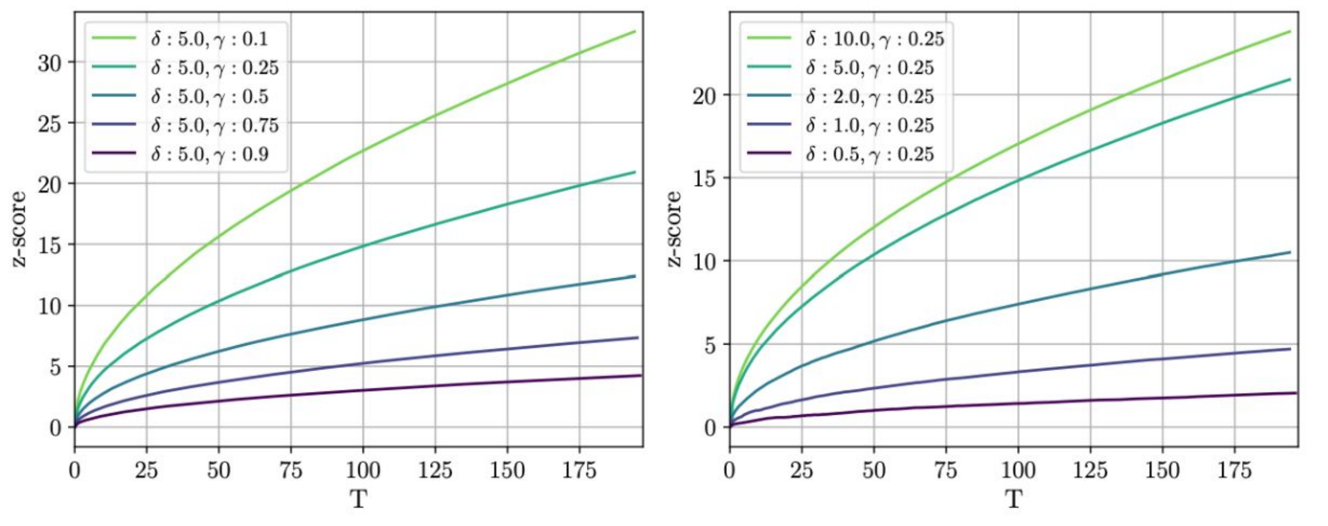
Hard Watermark

Soft Watermark



Row 1: Our Results

Row 2: Paper Results



Conclusion & Future Work

- ➔ Watermarked text can fall into a repeating loop, though strangely the paper does not seem to have this problem
- ➔ In general, it seems like watermarking does have a negative effect on text quality
- ➔ We would like to implement more of the attacks
- ➔ We would like to implement the private key algorithm
- ➔ Try this on bigger/better models

References

[1] Kirchenbauer, John, et al. "A watermark for large language models." *International Conference on Machine Learning*. PMLR, 2023